
Making the Most of What We Know: Towards Effective Use of Genomics Data

Terence Critchlow

*Center for Applied Scientific Computing
Lawrence Livermore National Laboratory
www.llnl.gov/CASC/people/critchlow*

PITTCON 2001

March 6, 2001

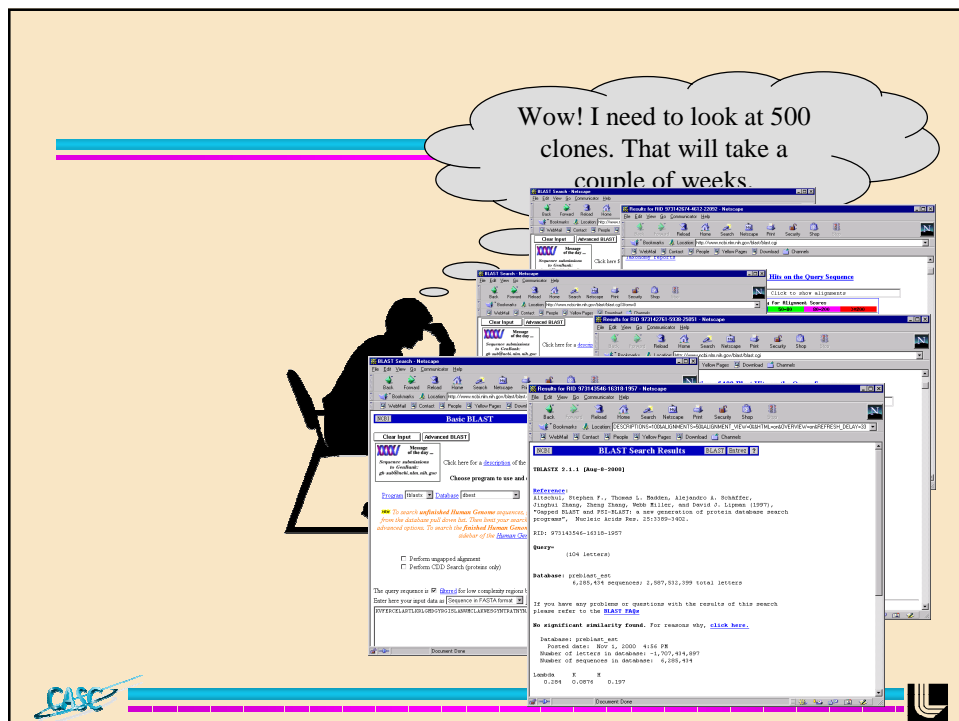
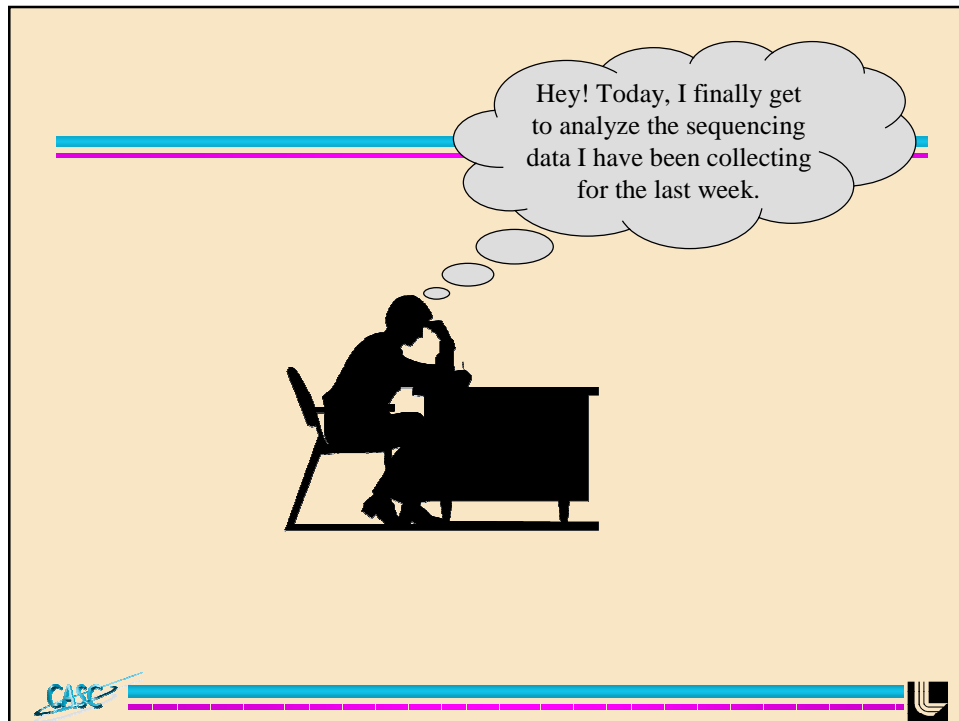
UCRL-VG-141500



Outline

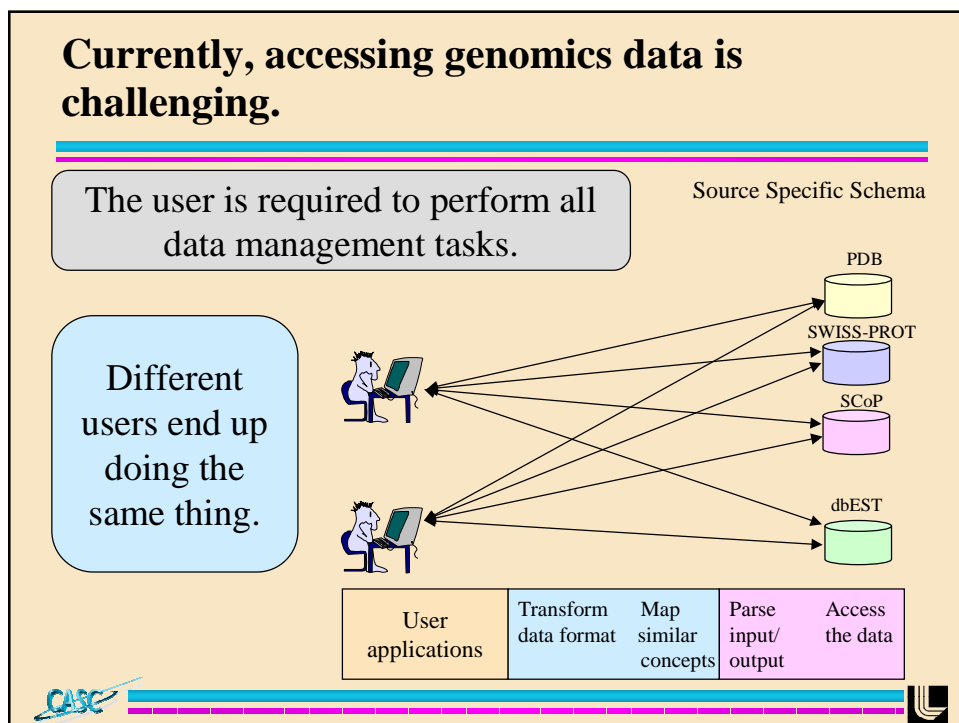
- **A day in the life of a geneticist/biologist.**
- **The business approach. (present)**
- **The impact of the web. (future)**
- **The key to success.**





Man, this sure would be easier if I could move between these different sites without having to reformat the data every time.

The collage shows four different web interfaces for protein structure and sequence data. The top left shows 'Structure Explorer - 134L' with a list of protein entries. The top right shows 'Structure Explorer - 133L' with a list of protein entries. The bottom left shows 'Structure Explorer - 137P' with a list of protein entries. The bottom right shows a 'view of SWISS-PROT: P79179' with a list of protein entries. A thought bubble from a person's head points to the screenshots, containing the text: 'Man, this sure would be easier if I could move between these different sites without having to reformat the data every time.'



The result of the current environment:

Scientists limit their searches to a small number of sources they are familiar with.

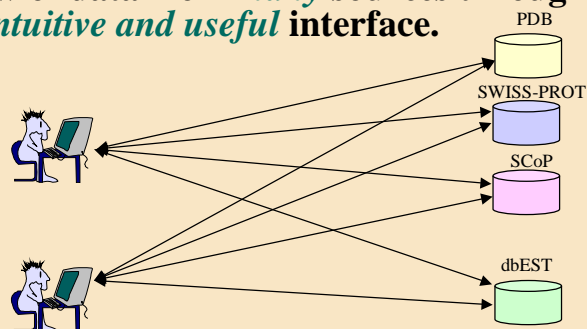
A huge amount of relevant information is not utilized.

Resources are wasted.
Time is wasted.
Knowledge is wasted.



What is our ideal environment?

A *single* location that provides *effective* access to a *consistent* view of data from *many* sources through an *intuitive and useful* interface.



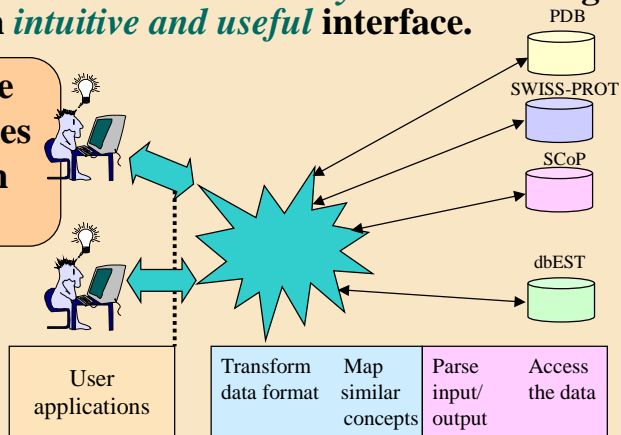
User applications	Transform data format	Map similar concepts	Parse input/output	Access the data
-------------------	-----------------------	----------------------	--------------------	-----------------



What is our ideal environment?

A *single* location that provides *effective* access to a *consistent* view of data from *many* sources through an *intuitive and useful* interface.

Businesses use data warehouses to accomplish this.



CASC

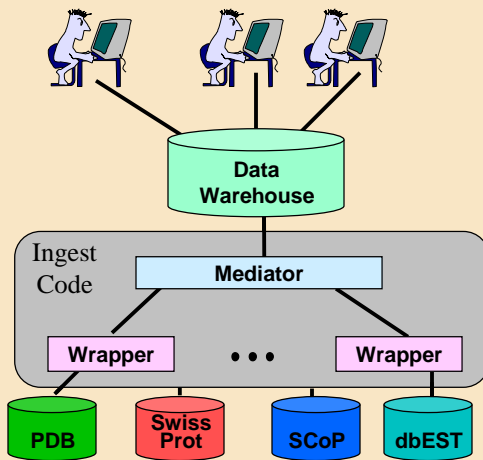


The business approach.

CASC



Data warehouses



- **Interfaces**

- provide intuitive access to the data
- possibly change data format to meet user expectations

- **Warehouse**

- stores a consistent view of data in a local repository

- **Mediator**

- transform data from source format to warehouse format

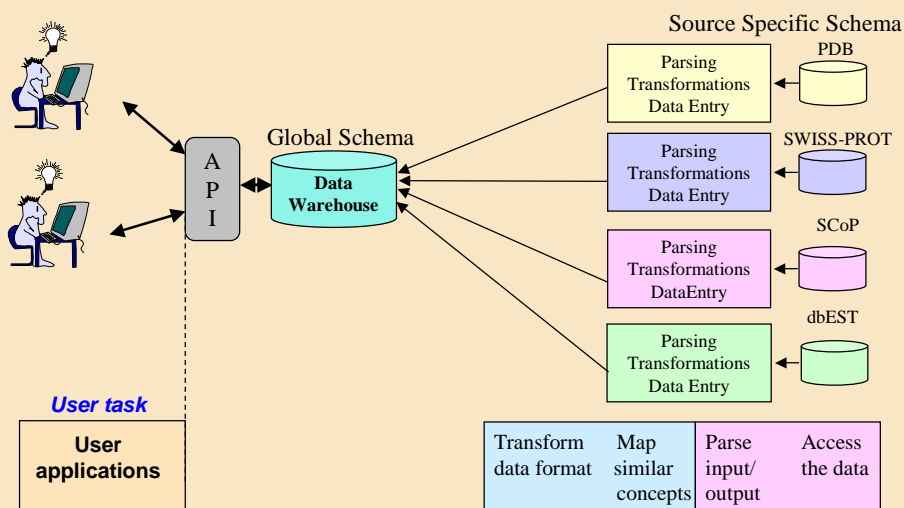
- **Wrappers**

- read data from source into internal representation

CASC



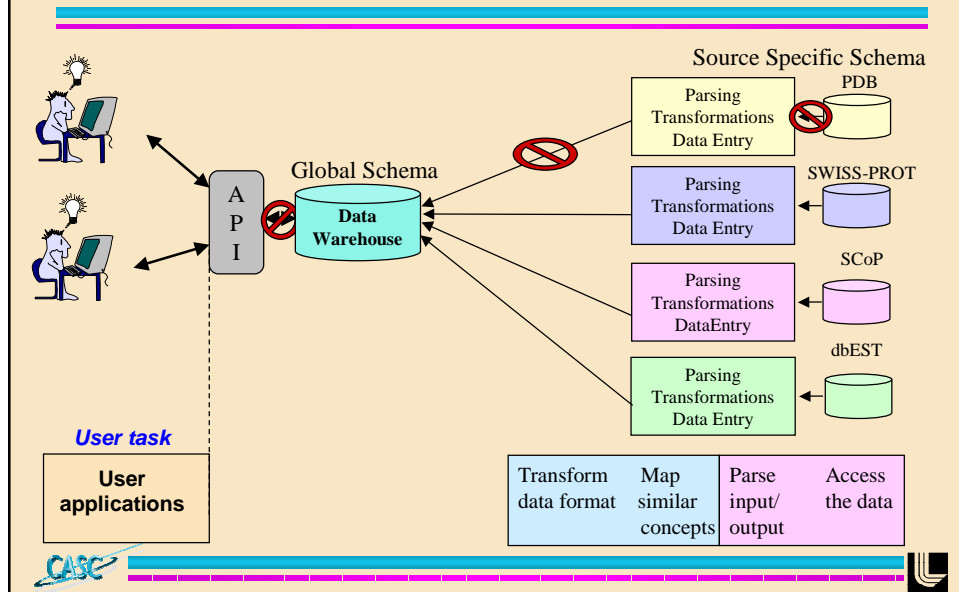
Typical approaches combine wrapper and mediator functionality.



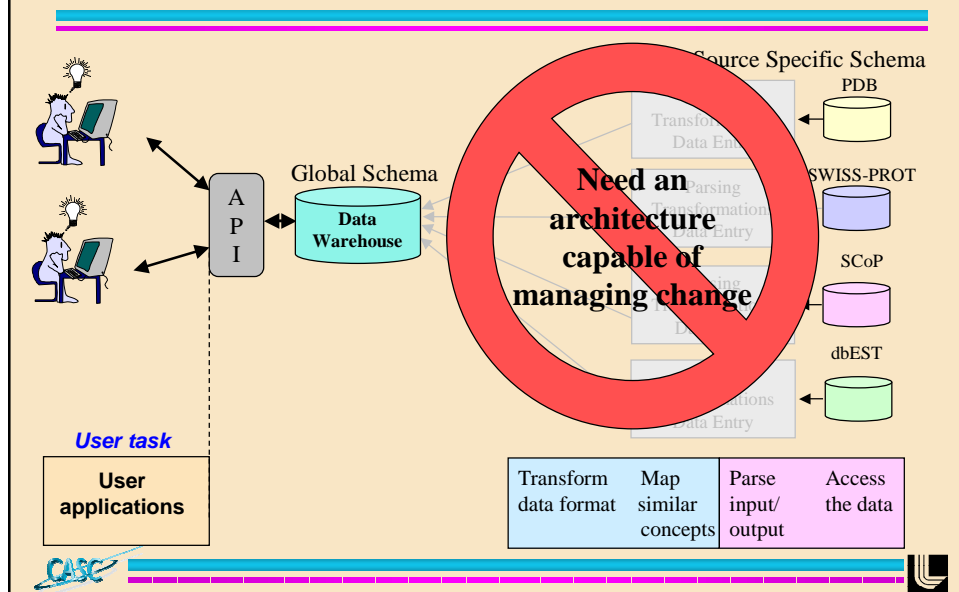
CASC



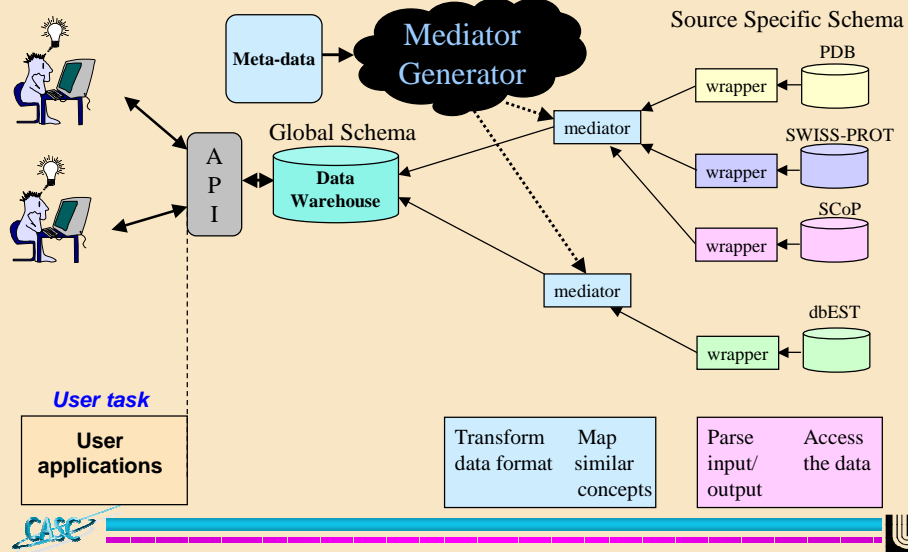
When a source changes, propagating the change can be time consuming.



When a source changes, propagating the change can be time consuming.

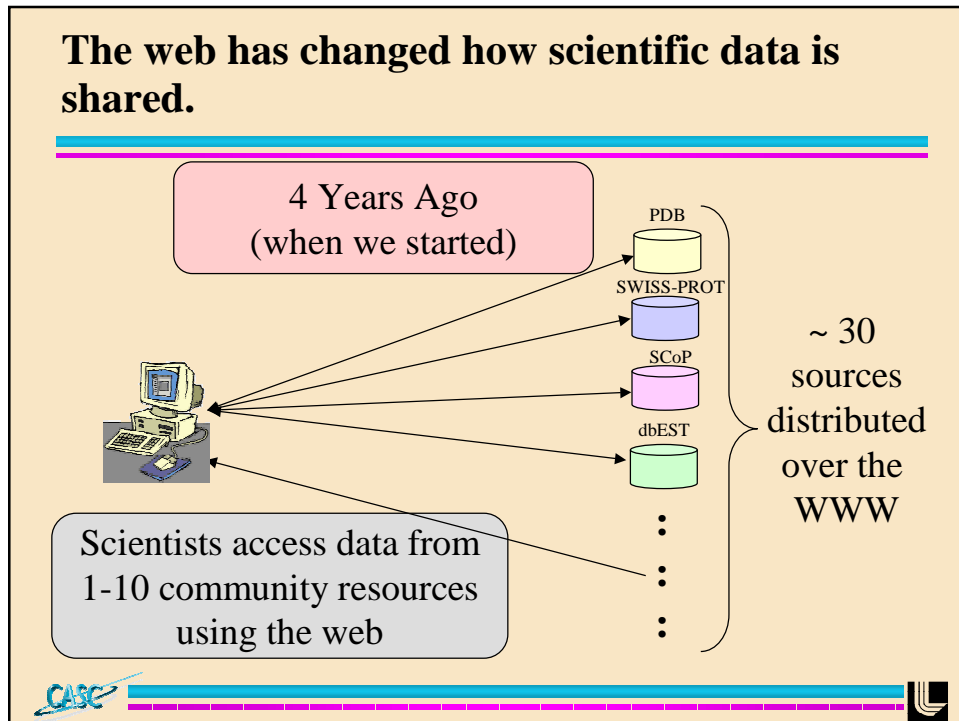


DataFoundry separates these tasks.

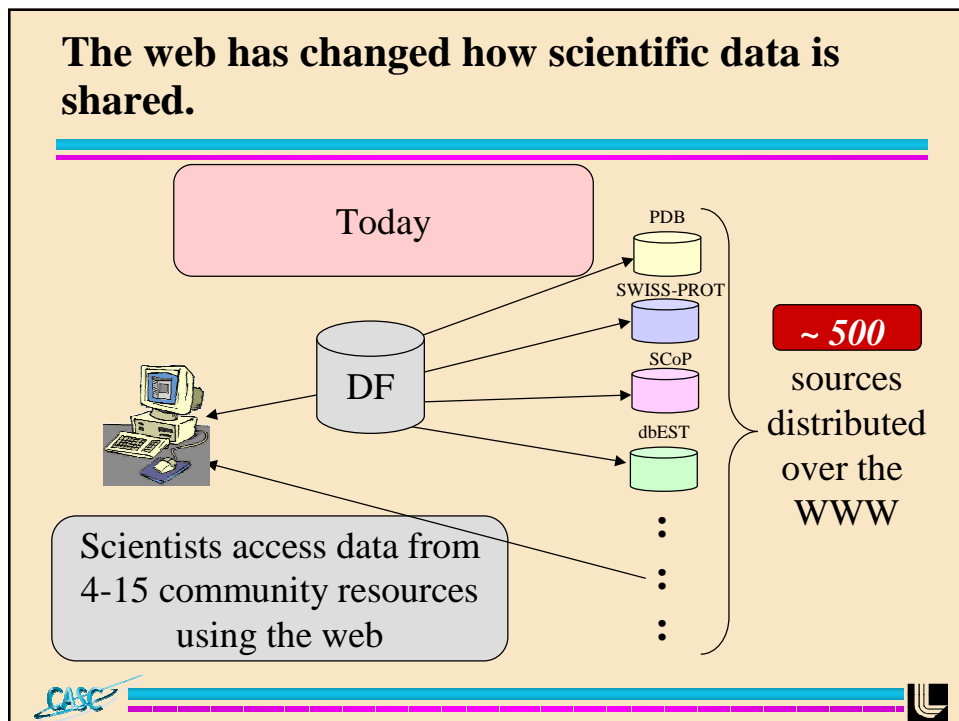


The impact of the Web.

The web has changed how scientific data is shared.



The web has changed how scientific data is shared.

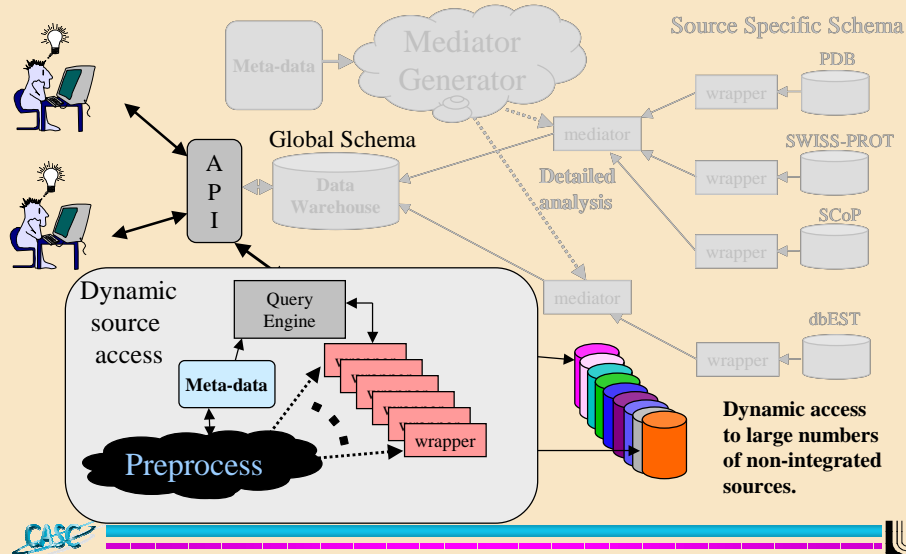


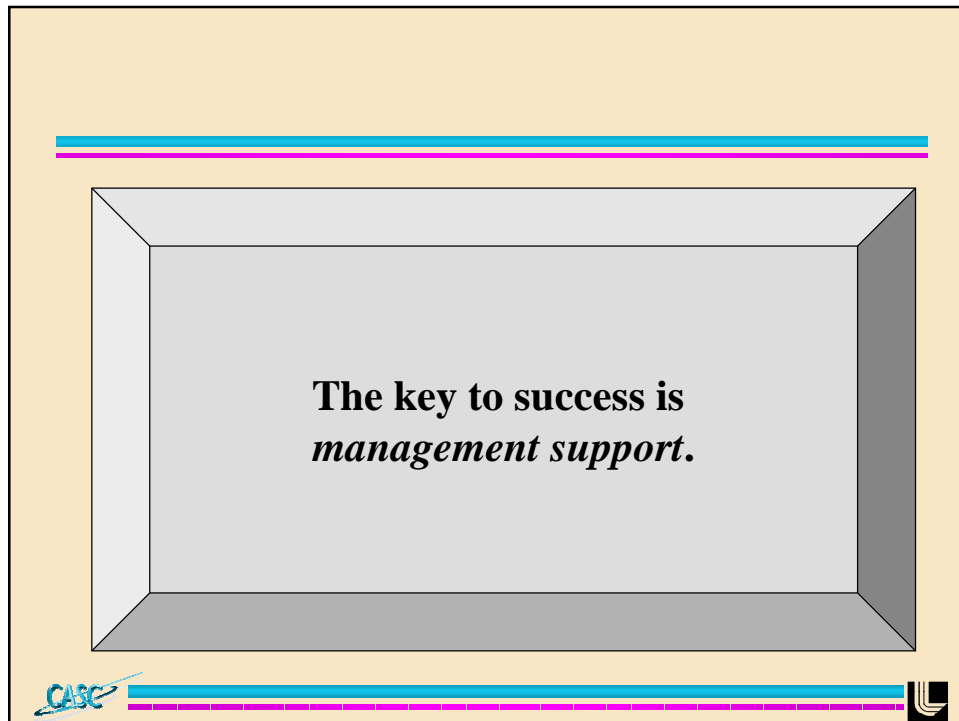
So, what makes this so much more difficult?

- **No central directory**
 - How do you find the sources in the first place?
- **Different data formats and semantics**
 - Once you find a source, how to you make sense of it?
- **Complex analysis**
 - Queries require more than simple data retrieval, they need to invoke complex programs.
- **Lots of data sources**
 - Too much work to be done manually.
 - Need to select appropriate subset for each user / query.
 - How do you present the results in an useful format?



A hybrid approach to information access holds the most promise.



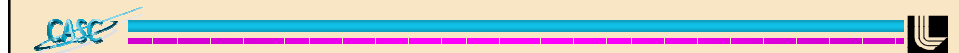


Ideally, management *fully* supports the bioinformatics effort.

Good management understands the complexity of the problems and provides the required resources.

- Schedules are reasonable
- Goals are realistic
- Staff is properly trained
- Funding is adequate and stable

The technical problems surrounding bioinformatics can be overcome.



Unfortunately, that is usually not the case.

Poor management underestimates the complexity of the problems and provides inadequate resources.

- Schedules are unreasonable
- Goals are unrealistic
- Morale is low & staff is poorly trained
- Funding is minimal and unstable

You are contributing to the problems facing bioinformatics instead of their solutions.



CASC



Conclusions

Bioinformatics has a long way to go before it can support scientific exploration.

The good news

The technical challenges surrounding bioinformatics can be overcome.

The bad news

The major challenge is political, not technical.

CASC



Conclusions

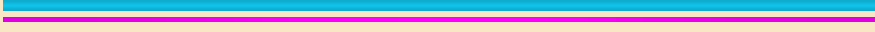
*Be part of the solution,
not part of the problem.*



Questions?

www.llnl.gov/CASC/people/critchlow





This work was performed under the auspices of the U.S.
Department of Energy by University of California Lawrence
Livermore National Laboratory under contract No. W-7405-
ENG-48.

